European Patent Office

# SUPPLEMENTARY EUROPEAN SEARCH REPORT

Application Number
EP 94 90 1238

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.5) |
|---|---|---|---|
| A | COMPUTER APPLICATIONS IN THE BIOSCIENCES, vol. 8, no. 5, October 1992 IRL PRESS, OXFORD, UK, pages 425-431, K.S. MAKAROVA ET AL. 'DIROM: an experimental design interactive system for directed mutagenesis and nucleic acids engineering' * the whole document * --- | 1-20 | C12N15/10 G06F15/42 |
| A | COMPUTER APPLICATIONS IN THE BIOSCIENCES, vol. 7, no. 4, October 1991 IRL PRESS, OXFORD, UK, pages 525-529, K. LUCAS ET AL. 'An improved microcomputer program for finding gene- or gene family-specific oligonucleotides suitable as primers for polymerase chain reaction or as a probe' * the whole document * --- | 1-20 | |
| A | COMPUTER APPLICATIONS IN THE BIOSCIENCES, vol. 8, no. 2, April 1992 IRL PRESS, OXFORD, UK, pages 121-127, P.A. PEVZNER 'Statistical distance between texts and filtration methods in sequence comparison' * the whole document * --- | 1-20 | TECHNICAL FIELDS SEARCHED (Int.Cl.5) C12N G06F |
| A | PROC. NATL.ACAD SCI., vol. 89, July 1992 NATL. ACAD SCI.,WASHINGTON,DC,US;, pages 6090-6093, M.S. WATERMAN ET AL. 'Parametric sequence comparisons' * the whole document * --- | 1-20 | |

-/--

The supplementary search report has been drawn up for the claims attached hereto.

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| THE HAGUE | 8 January 1996 | Hornig, H |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C04)

European Patent Office

# SUPPLEMENTARY EUROPEAN SEARCH REPORT

Application Number
EP 94 90 1238

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.5) |
|---|---|---|---|
| P,X | WO-A-93 15221 (HITACHI CHEMICAL CO LTD ;HITACHI CHEMICAL RESEARCH CENT (US)) 5 August 1993<br>* page 79, line 7 - page 102, line 17; claims 1,23-40 *<br>----- | 1-20 | |
| | | | TECHNICAL FIELDS SEARCHED (Int.Cl.5) |

The supplementary search report has been drawn up for the claims attached hereto.

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| THE HAGUE | 8 January 1996 | Hornig, H |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C04)

CLAIMS

94901238.9

1. A programmed computer system for designing optimal oligonucleotide sequences for use with a gene sequence data source comprising:

first input means for introducing user-selected gene sequence into the computer system;

memory means for storing user-selected gene sequence;

means for accessing gene sequence data from said gene sequence data source;

means for performing exact and inexact match modelling between gene sequences;

means for performing hybridization strength modelling on gene sequences;

means for selecting either of said modelling means; and,

means for presenting the results of said modelling to present candidate oligonucleotide sequences.

2. A programmed computer system in accordance with claim 1, wherein said means for performing exact and inexact match modelling utilizes said accessing means to introduce a user-selected set of gene sequence data and user-selected set of target gene sequence data from said gene sequence data source into the computer system and said memory means to store said gene sequence data and said target gene sequence data and wherein said means for performing exact and inexact match modelling includes:

means for determining a minimum sequence length;

means for creating a look-up hash table and linked list in memory for each gene sequence in said gene sequence data and each of said garget gene sequences;

means for calculating the minimum length of any matching gene subsequence of said gene sequence data and said target gene sequence data;

means for comparing each base pair character in each said target sequence stored in a hash table in memory to each base pair character of said gene sequence stored in a hash table in memory;

means for finding a matching seed by determining if the said comparison results in a matching gene subsequence of length equal to said calculated minimum length;

means for comparing base pair characters behind and ahead of said seed to determine if there exists an extended match of a subsequence of base pair characters of length greater than the calculated minimum length, resulting in a current hit sequence;

means for calculating whether said current hit sequence is longer than said minimum sequence length, resulting in a current candidate oligonucleotide sequence;

means for storing said current candidate oligonucleotide sequence; and,

wherein said presenting means provides said current candidate oligonucleotide sequence to the user.

3. A programmed computer system in accordance with claim 2, wherein said computer system includes:

means for calculating the melting temperature for each candidate oligonucleotide sequence;

means for tracking the number and melting temperature of the matches for each candidate oligonucleotide sequence;

means for tracking the location of a set number of the best candidate oligonucleotide sequences preferably employing a priority queue by sorting said candidate oligonucleotide sequence in reverse order and sorting said oligonucleotides by hybridization strength; and,

wherein said presenting means is operative to present said additional results to the user; and,

wherein said presenting means operably provides said melting temperature to the user.

4. A programmed computer system in accordance with claims 2 or 3, wherein said first input means in operative to introduce a user-selected maximum number of mismatches and a user-selected minimum candidate oligonucleotide sequence length into the computer system, and wherein said means for calculating the minimum length of any matching gene subsequence of said gene sequence data and said target gene sequence data comprises the steps of:

means for subtracting said maximum number of mismatches from said minimum candidate oligonucleotide sequence length to give a first result;

means for dividing said first result by said maximum number of mismatches plus one to give a second result;

means for incrementing said second result by one if the remainder is not equal to zero to give a third result; and,

means for truncating said third result to an integer, and preferably said means for calculating the hairpin characteristics of said candidate oligonucleotide sequence comprises the steps of:

calculating a complementary sequence to the candidate oligonucleotide sequence by reversing the base pair order of the candidate oligonucleotide and substituting complementary base pairs;

comparing each character of said original candidate oligonucleotide sequence and said complementary sequence;

finding the longest match between said original candidate oligonucleotide sequence and said complementary sequence; and,

saving the match with the longest hairpin distance if any two matches have the same length;

means for storing hairpin characteristics; and,

wherein said presenting means provides said hairpin characteristics to the user.

5. A programmed computer system in accordance with any preceding claim, wherein said means for performing hybridization strength modelling utilizes said first input means to introduce a user-selected screening threshold into the computer system and said accessing means to introduce a user-selected set of gene sequence data from said gene sequence data source into the computer system, and said memory means to store said gene sequence data, said target gene sequence data and said screening threshold and wherein said means for performing hybridization strength modelling comprises:

means for preprocessing said target gene sequence data and said gene sequence data by selecting only those sequences without introns;

means for forming a preparation file of gene sequence

fragments by cutting said target gene sequences into fixed length target gene subsequences and sorting said subsequences in lexicographical order;

means for merge sorting said gene sequences;

means for forming multiple lists of screens by forming lists of subsequences of the preparation file of length equal to said screening threshold;

means for indexing, sorting and storing said screens in said memory means;

means for sequentially comparing said preparation file gene sequences with each of said screens to design candidate oligonucleotide sequences;

means for calculating the hybridization strengths between a gene sequence and all candidate oligonucleotide sequences containing that gene sequence by accounting for Guanine-Cytosine (GC) and Adenine-Thymine (AT) base pair content of the gene sequence and the number of mismatches between said preparation file sequences and a said screen when said comparison results in a match;

means for preparing the candidate oligonucleotide sequence and hybridization strength for presentation to the user; and

wherein said presenting means provides the candidate oligonucleotide sequence and hybridization strength to the user.

6. A programmed computer system in accordance with claim 5, wherein said computer system includes:

means for assigning a numerical score to each said gene

sequence; and

means for sorting said gene sequences in accordance with said numerical score, and/or preferably wherein means for assigning a numerical score to each said gene sequence operates by tallying the quantity "exp" where "exp" = $e^{-Tm}$ and wherein Tm is the melting temperature for the said gene sequence; and

means for sorting said gene sequences in accordance with said numerical score.

7. A programmed computer system in accordance with claim 5, wherein said means for calculating the hybridization strengths between a gene sequence and all candidate oligonucleotide sequences containing that gene sequence comprises the steps of:

accessing gene sequence data from said gene sequence data source;

comparing base pairs of a first gene sequence and a second gene sequence to determine if a match exists;

incrementing said first gene sequence's bound strength by some first number if a base pair character in said first gene sequence and said second gene sequence match and the matched base pair is equal to a combination of the bases Guanine (G) and Cytosine (C);

incrementing said first gene sequence's bound strength by some second number if a base pair character in said first gene sequence and said second gene sequence match and the matched base pair is equal to a combination of the bases Adenine (A)

and Thymine (T);

decrementing said first gene sequence's bound strength by a third number if there is no match in base pairs between said first gene sequence and said second gene sequence;

comparing said first gene sequence's bound strength to said first gene sequence's unbound strength;

setting said first gene sequence's unbound strength equal to its bound strength if said first gene sequence's bound strength is greater than said first gene sequence's unbound strength; and

resetting said first gene sequence's bound strength to zero if said first gene sequence's unbound strength is less than zero.

8. A programmed computer system in accordance with claim 5, wherein said computer system includes a means for calculating the hairpin characteristics of said candidate oligonucleotide sequence;

means for preparing the hairpin characteristics for presentation to the user; and,

wherein said presenting means provides the hairpin characteristics to the user, wherein preferably means for calculating the hair pin characteristics of said candidate oligonucleotide sequence comprises the steps of:

calculating a complementary sequence to the candidate oligonucleotide sequence by reversing the base pair order of the candidate oligonucleotide sequence and substituting complementary base pairs;

comparing each character of said original candidate

oligonucleotide sequence and said complementary sequence;

finding the longest match between said original candidate oligonucleotide sequence and said complementary sequence; and

saving the match with the longest hairpin distance if any two matches have the same length;

means for preparing the hairpin characteristics for presentation to the user; and

wherein said presenting means provides the hairpin characteristics to the user.

9. A programmed computer system in accordance with claim 1, wherein said means for performing exact and inexact match modelling utilizes said accessing means to introduce a user-selected set of gene sequence data and user-selected set of target gene sequence data from said gene sequence data source into the computer system and said memory means to store said gene sequence data and said target gene sequence data and wherein said means for performing exact and inexact match modelling includes:

means for determining a minimum sequence length;

means for creating a look-up hash table and linked list in memory for each gene sequence in said gene sequence data and each of said target gene sequences;

means for calculating the minimum length of any matching gene subsequence of said gene sequence data and said target gene sequence data;

means for transforming base characters in each said target sequence and in each said gene sequence into numeric digits;

means for comparing each base pair digit in each said target sequence stored in a hash table in memory to each base pair digit of said gene sequence stored in a hash table in memory;

means for finding a matching seed by determining if the said comparison results in a matching gene subsequence of length equal to said calculated minimum length;

means for comparing base pair digits behind and ahead of said seed to determine if there exists an extended match of a subsequence of base pair digits of length greater than the calculated minimum length, resulting in a current hit sequence;

means for calculating whether said current hit sequence is longer than said minimum sequence length, resulting in a current candidate oligonucleotide sequence;

means for storing said current candidate oligonucleotide sequence; and

wherein said presenting means provides said current candidate oligonucleotide sequence to the user.

10. A programmed computer system for designing candidate oligonucleotide sequences for use with a gene sequence data source including:

first input means for introducing user-selected gene sequence, design, model and presentation criteria and user-specified sequence length into the computer system;

memory means for storing said gene sequence, design, model and presentation criteria and said sequence length;

means for accessing gene sequence data from said gene

sequence data source;

wherein said accessing means is operative to introduce a user-selected set of gene sequence data and a user-selected set of target gene sequence data from said gene sequence data source into the computer system;

wherein said criteria are used for comparison of gene sequence data and target gene sequence data;

means for comparing said gene sequences against said target gene sequences employing said criteria;

means for calculating candidate oligonucleotide sequences of said sequence length that are either common to a pool of user-specified gene sequences or specific to a particular user-specified gene sequence;

means for calculating the homology between the candidate oligonucleotide sequences and said gene sequence data;

means for calculating a candidate oligonucleotide sequence's hairpin characteristics;

means for displaying in multiple dimensions the gene sequences which result from the comparisons and calculations characterized in that said display format exhibits:

the starting position of each candidate oligonucleotide sequence in one dimension;

a candidate oligonucleotide sequence's specificity to the target gene sequence in a second dimension; and

superimposed melting temperatures of gene sequences in contrasting presentations in at least an apparent third dimension;

wherein said display further includes a cursor movable

along one dimension of said display that selects a position for an expansion of data representing the homology between the candidate oligonucleotide sequences and said gene sequence data; and

wherein said display is operative to display in alphanumeric form the homology between the candidate oligonucleotide sequences and said gene sequence data; and

wherein said display is operative to provide an expansion of data including presenting

false hybridizations at various melting temperatures for all candidate oligonucleotide sequences;

the location of each false hybridization;

a candidate oligonucleotide sequence's starting position; and,

hairpin characteristics of each candidate oligonucleotide sequence.

11. A method for designing candidate oligonucleotide sequences by performing exact and inexact match modelling for use with a gene sequence data source comprising the steps of:

introducing user-selected gene sequence into a computer system;

accessing gene sequence data from said gene sequence data source;

storing user-selected gene sequence in the memory of the computer system;

accessing the gene sequence source to introduce the user-selected set of gene sequence data and user-selected set of target gene sequence data from said gene sequence data source

into the computer system;

storing said gene sequence data and said target gene sequence data in the memory of the computer system;

determining a minimum sequence length;

creating a look-up hash table and linked list in memory for each gene sequence in said gene sequence data and each of said target gene sequences;

calculating the minimum length of any matching gene subsequence of said gene sequence data and said target gene sequence data;

comparing each base pair character in each said target sequence stored in a hash table in memory to each base pair character of said gene sequence stored in a hash table in memory;

determining a matching seed by determining if the said comparison results in a matching gene subsequence of length equal to said calculated minimum length;

comparing base pair characters behind and ahead of said seed to determine if there exists an extended match of a subsequence of base pair characters of length greater than the calculated minimum length, resulting in a current hit sequence;

calculating whether said current hit sequence is longer than said minimum sequence length, resulting in a current candidate oligonucleotide sequence;

storing said current candidate oligonucleotide sequence in the memory of the computer system; and,

presenting a representation of said current candidate

oligonucleotide sequence to the user.

12. A method in accordance with claim 11, wherein said method includes the steps for performing additional calculations for each candidate oligonucleotide probe, said additional calculations comprising:

calculating the melting temperature for each candidate oligonucleotide sequence;

tracking the number and melting temperature of the matches for each candidate oligonucleotide sequence;

tracking the location of a set number of the best candidate oligonucleotide sequences preferably employing a priority queue by sorting said candidate oligonucleotide sequences in reverse order and sorting said candidate oligonucleotide sequences by hybridization strength; and

presenting said additional results to the user.

13. A method in accordance with claim 12, wherein said step for calculating the minimum length of any matching gene subsequence comprises:

introducing a user-selected maximum number of mismatches and a user-selected minimum candidate oligonucleotide sequence length into the computer system;

subtracting said maximum number of mismatches from said minimum candidate oligonucleotide sequence length to give a first result;

dividing said first result by said maximum number of mismatches plus one to give a second result;

incrementing said second result by one if the remainder is not equal to zero to give a third result; and,

truncating said third result to an integer.

14. A method in accordance with any one of claims 11 to 13, wherein said method includes the step of calculating the hairpin characteristics of said candidate oligonucleotide sequence, preferably said method includes the step of calculating the hairpin characteristics of said candidate oligonucleotide sequence comprising:

calculating a complementary sequence to the candidate oligonucleotide sequence by reversing the base pair order of the candidate oligonucleotide sequence and substituting complementary base pairs;

comparing each character of said original candidate oligonucleotide sequence and said complementary sequence;

finding the longest match between said original candidate oligonucleotide sequence and said complementary sequence; and,

saving the match with the longest hairpin distance if any two matches have the same length.

15. A method for designing candidate oligonucleotide sequences by performing hybridization strength modelling for use with a gene sequence data source comprising the steps of:

introducing user-selected gene sequence and a user-selected screening threshold into a computer system;

storing user-selected gene sequence and said screening threshold in the memory of the computer system;

accessing the gene sequence source to introduce the user-selected set of gene sequence data and a user-selected set of target gene sequence data from said gene sequence data source into the computer system;

storing said gene sequence data and said target gene sequence data in the memory of the computer system;

preprocessing said target gene sequence data and said gene sequence data by selecting only those sequences without introns;

forming a preparation file of gene sequence fragments by cutting said target gene sequences into fixed length target gene subsequences and sorting said subsequences in lexicographical order;

merge sorting said gene sequences;

forming multiple lists of screens by forming lists of subsequences of the preparation file of length equal to said screening threshold;

indexing and sorting said screens in memory;

storing said screens in the memory of the computer system;

sequentially comparing said preparation file gene sequences with each of said screens to design candidate oligonucleotide sequences;

calculating the hybridization strengths between a gene sequence and all candidate oligonucleotide sequences containing that gene sequence by accounting for Guanine-Cytosine (GC) and Adenine-Thymine (AT) base pair content of the gene sequence and the number of mismatches between said preparation file sequences and a said screen when said comparison results in a match;

preparing the candidate oligonucleotide sequence and hybridization strength for presentation to the user; and,

presenting the candidate oligonucleotide sequence and hybridization strength to the user, and preferably wherein said method includes the steps for performing additional calculations for each candidate oligonucleotide probe, said additional calculations comprising:

calculating the melting temperature for each candidate oligonucleotide sequence;

tracking the number and melting temperature for each candidate oligonucleotide sequence;

tracking the location of a set number of the best candidate oligonucleotide sequences preferably employing a priority queue by sorting said candidate oligonucleotide sequences in reverse order and sorting said candidate oligonucleotide sequences by hybridization strength; and

presenting said additional results to the user.

16. A method in accordance with claim 15, to use with a gene sequence data source, programmed to determine hybridization strength comprising the steps of:

comparing base pairs of a first gene sequence and a second gene sequence to determine if a match exists;

incrementing said first gene sequence's bound strength by some first number if a base pair character in said first gene sequence and said second gene sequence match and the matched base pair is equal to a combination of the bases Guanine (G) and Cytosine (C);

incrementing said first gene sequence's bound strength by some second number if a base pair character in said first gene sequence and said second gene sequence match and the matched

base pair is equal to a combination of the bases Adenine (A) and Thymine (T);

decrementing said first gene sequence's bound strength by a third number if there is no match in base pairs between said first gene sequence and said second gene sequence;

comparing said first gene sequence's bound strength to said first gene sequence's unbound strength;

setting said first gene sequence's unbound strength equal to its bound strength if said first gene sequence's bound strength is greater than said first gene sequence's unbound strength; and

resetting said first gene sequence's bound strength to zero if said first gene sequence's unbound strength is less than zero; and

preferably wherein said first and second numbers are greater than zero; and/or

preferably wherein said second number is in the order of 42% of said first number; and/or

wherein said third number is in the order of 5% larger than said first number; and/or

preferably wherein said method includes the step of calculating the hairpin characteristics of said candidate oligonucleotide sequence, wherein preferably including the steps of:

calculating a complementary sequence to the candidate oligonucleotide sequence by reversing the base pair order of the candidate oligonucleotide sequence and substituting complementary base pairs;

comparing each character of said original candidate oligonucleotide sequence and said complementary sequence;

finding the longest match between said original candidate oligonucleotide sequence and said complementary sequence; and,

saving the match with the longest hairpin distance if any two matches have the same length.

17. A method for designing candidate oligonucleotide sequences for use with a gene sequence data source comprising the steps of:

introducing user-selected gene sequence and user-specified sequence length into a computer system;

storing said gene sequence and said sequence length in the memory of the computer system;

accessing gene sequence data from said gene sequence data source;

accessing the gene sequence source to introduce the user-selected set of gene sequence data and a user-selected set of target gene sequence data from said gene sequence data source into the computer system;

comparing said gene sequences against said target gene sequences employing said criteria;

calculating candidate oligonucleotide sequences of said sequence length that are either common to a pool of user-specified gene sequences or specific to a particular user-specified gene sequence;

calculating the homology between the candidate oligonucleotide sequences and said gene sequence data;

displaying in multiple dimensions the gene sequences

which result from the comparisons and calculations characterized in that said display format exhibits;

the starting position of each candidate oligonucleotide sequence in one dimension;

a candidate oligonucleotide sequence's specificity to the target gene sequence in a second dimension; and,

superimposed melting temperatures of gene sequences in contrasting presentations in at least an apparent third dimension, and wherein preferably said method includes the step of calculating a candidate oligonucleotide sequence's hairpin characteristics, preferably comprising the steps of:

calculating a complementary sequence to the said gene sequence by reversing the base pair order of the gene sequence and substituting complementary base pairs;

comparing each character of said original gene sequence and said complementary sequence;

finding the longest match between said original gene sequence and said complementary sequence; and,

saving the match with the longest hairpin distance if any two matches have the same length.

18. A method to determine hybridization strength between two or more gene sequences for use with a gene sequence data source, comprising the steps of:

accessing gene sequence data from said gene sequence data source;

comparing base pairs of a first gene sequence and a second gene sequence to determine if a match exists;

incrementing said first gene sequence's bound strength by

some first number if a base pair character in said first gene sequence and said second gene sequence match and the matched base pair is equal to a combination of the bases Guanine (G) and Cytosine (C);

incrementing said first gene sequence's bound strength by some second number if a base pair character in said first gene sequence and said second gene sequence match and the matched base pair is equal to a combination of the bases Adenine (A) and Thymine (T);

decrementing said first gene sequence's bound strength by a third number if there is no match in base pairs between said first gene sequence and second gene sequence;

comparing said first gene sequence's bound strength to said first gene sequence's unbound strength;

setting said first gene sequence's unbound strength equal to its bound strength if said first gene sequence's bound strength is greater than said first gene sequence's unbound strength; and

resetting said first gene sequence's bound strength to zero if said first gene sequence's unbound strength is less than zero, and wherein preferably said first and second numbers are greater than zero, and/or preferably wherein said second number is in the order of 42% of said first number, and/or wherein said third number is in the order of 5% larger than said first number.

19. A method of calculating the minimum length of any matching gene subsequence comprising:

introducing a user-selected maximum number of mismatches

and a user-selected minimum candidate oligonucleotide sequence length;

subtracting said maximum number of mismatches from said minimum candidate oligonucleotide sequence length to give a first result;

dividing said first result by said maximum number of mismatches plus one to give a second result;

incrementing said second result by one if the remainder is not equal to zero to give a third result; and,

truncating said third result to an integer.

20. A method of calculating hairpin characteristics for a gene sequence comprising:

calculating a complementary sequence to the said gene sequence by reversing the base pair order of the gene sequence and substituting complementary base pairs;

comparing each character of said original gene sequence and said complementary sequence;

finding the longest match between said original gene sequence and said complementary sequence; and,

saving the match with the longest hairpin distance if any two matches have the same length.